A Novel Ant based Clustering of Gene Expression Data using MapReduce Framework

Bhavani R, Assistant Professor
Department of CSE
Government College of Technology
Coimbatore, India.
bhavanirajasekar@gmail.com

Dr.G.Sudha Sadasivam, Professor
Department of CSE
PSG College of Technology
Coimbatore, India.
sudhasadhasivam@yahoo.com

Abstract— Genes which exhibit similar patterns are often functionally related. Microarray technology provides a unique tool to examine how a cell's gene expression pattern changes under various conditions. Analyzing and interpreting these gene expression data is a challenging task. Clustering is one of the useful and popular methods to extract useful patterns from these gene expression data. In this paper multi colony ant based clustering approach is proposed. The whole processing procedure is divided into two parts: The first is the construction of Minimum spanning tree from the gene expression data using MapReduce version of ant colony optimization techniques. The second part is clustering, which is done by cutting the costlier edges from the minimum spanning tree, followed by one step k-means clustering procedure. Applied to different file sizes of gene expression data over different number of processors, the proposed approach exhibits good scalability and accuracy.

Keywords- *Bioinformatics, Gene expression data, Multi colony ant system, Data mining, Clustering, MapReduce programming)*

I. INTRODUCTION

Data Mining is the process of analyzing large datasets to find useful patterns. Microarray technology is an experimental technique that can measure expression levels of hundreds and thousands of genes simultaneously. Analysis of gene expression data involves many computational tools for searching genes of interest, clustering and classification to find meaningful interpretation from huge volume of data. This paper aims at clustering the genes in the gene expression data. It helps in understanding gene functions and regulatory networks and assists in the diagnostics of disease conditions and effects of medical treatment.

Clustering is one of the important methods in the field of data mining which aims at grouping objects into clusters such that the objects from the same cluster are similar and objects from different clusters are dissimilar. The similarity measurement is calculated through distance function. It is an unsupervised learning technique where the given dataset is analyzed and grouped into meaningful clusters without the prior knowledge of the classes in the dataset [1].

Traditional clustering algorithms can be broadly classified into two categories namely partitioning method and hierarchical methods. K-means clustering is a partitioning method of clustering which partitions the given dataset into k clusters. It is one of the easy and efficient methods for clustering and the parameter k is crucial. Clustering based on metaheuristic algorithms is emerging as an alternative to more conventional clustering techniques.

In this paper, an ant based metaheuristic algorithm is proposed to perform gene expression data clustering. Ant colony optimization (ACO) is a kind of metaheuristic based on the behaviour of ants seeking a path between their colony and a source of food. Solutions for a given problem are constructed by random walks of artificial ants on a so-called construction graph, which has pheromone (weights) on the edges. Some of the problems in conventional clustering methods like clusters with arbitrary shapes, clusters with outliers are resolved using ACO based clustering. Since the work involves processing huge size of data, that is computationally-intensive and time-consuming, a MapReduce model for clustering is proposed. MapReduce programming model is typically used in distributed computing on clusters of computers. The model abstract distributed computing in two steps. The Map step is applied on the input data and produces a list of intermediate results. The Reduce step is applied to the intermediate results to perform some kind of merging operation to produce the output. Developers need to code Map and Reduce functions, and then submit the job to the MapReduce operating environment. Hadoop is open-source implementation of MapReduce computing model [2].

II. RELATED WORK

Study of related literature can be grouped under two categories namely ant based clustering methods and parallelism of ACO algorithm.

The sum of k-nearest neighbor distances metric and a shrinking range strategy is accommodated with ant colony optimization algorithm to resolve the problem of clusters with arbitrary shapes, clusters with outliers and bridges between outliers [3]. ACO based feature selection for image clustering is adopted in[4] and is used in content based image retrieval. Preprocessing of input to k-means clustering is done using ant based self organizing maps (SOM). It embeds the exploitation and exploration rules of state transition into conventional SOM algorithm [5]. The next position of the particle in the PSO algorithm is found using ACO and is taken as the initial clusters of the k-means approach [6]. ACO with different flavor (ACODF) uses simulated annealing concept for ants to decreasingly visit the amount of cities and uses tournament selection strategy to choose a path [7]. The result of k-means clustering is taken as the elicitation information of ACO [8]. Shortest path between the documents are found using ACO in the first phase. Group of documents which are alike are separated in the second phase [9]. All the above approaches used single ant system and are hybridized with other machine learning algorithms.

Parallel implementation of ant colony optimization algorithm is adopted for industrial scheduling problem using OpenMP implementation on shared memory architecture [10]. Multi colony distributed ACO approach is used in solving HP protein folding problem in both two and three dimensions. The algorithm was run on IBM Blade centre comprising of 9 nodes [11]. Parallelization is achieved using GPUs for solving travelling sales person problem using ant colony optimization [12]. Message Passing Interface based parallel ant colony optimization is adopted to solve travelling salesman problem [13]. MapReduce model of ACO is formulated to solve combinatorial optimization problems [14]. Max-min ant system is parallelized using MapReduce programming model to solve travelling salesman problem [15]. ACO is adopted in changing the activation threshold in the MapReduce model of DE-ACO-kmeans clustering [16].

In all the above approaches ACO is adopted to solve travelling salesman problem and the same was parallelized using OpenMP, MPI, Grid environment, GPU cards and MapReduce. The proposed approach uses MapReduce model of multi colony ant system to construct minimum spanning tree in the first phase. The costlier edges were removed to form clusters in the second phase.

III. Proposed Approach

The proposed approach uses a multi colony ant system for clustering the gene expression data. As a preprocessing step, the construction graph is built from the gene expression data. Each node of the construction graph represents a gene in the gene expression data. The ant colonies perform random walk on the construction graph to construct the solution.

A. *Basic ACO operations*

Given a construction graph, the MST is the spanning tree with minimum cost and contains all the nodes in the graph without forming any cycle. [17] The probability of choosing the next node j by ant k occupying node i is given by

$$p_{ij} = \frac{(\tau_{ij})^{\alpha} (\eta_{ij})^{\beta}}{\sum_{k \in S} (\tau_{ik})^{\alpha} (\eta_{ik})^{\beta}} \tag{1}$$

where S represents the list of nodes not visited by ant k. τ_{ij} represents the amount of pheromone trail between node i,j. α and β are the control parameters that control the relative importance of trial versus visibility. $\eta_{ij} = 1/d_{ij}$ where d_{ij} is the Euclidean distance between node i and node j. Initially $\tau_{ij} = 1/n$ where n is the number of nodes in the construction graph. After all the ants complete their tour, the pheromone is evaporated on all the arcs. Pheromone evaporation is implemented by

$$\tau_{ij} = \rho \cdot \tau_{ij} \tag{2}$$

where ρ is the decay constant such that $\rho < 1$. The new amount of pheromone is deposited only in the path visited by the best ant. Pheromone deposition is given by

$$\tau_{ij} = \tau_{ij} + \frac{Q}{M} \tag{3}$$

where M is the cost of the best-so-far path and Q is a constant.

B. *Clustering using Multi colony ant system*

Ant algorithms are good candidates for parallelization using MapReduce framework. In MapReduce model of multi colony ant system, every mapper holds a colony of ant and builds a MST. After each generation the colonies exchange information about their solutions. The best-so-far solution is found by the reducer which aggregates all the values from the mapper. The computations for the new pheromone information are also done by the reducer and are distributed to all the map function for next generation. The algorithm is repeated until some stopping criterion is met or the best found solution did not change for generations. Finally, the threshold value is found from the weights of all the edges in the minimum spanning tree using

$$Th = Q_3 - Q_1 \tag{4}$$

where Th is the threshold value, Q_1 and Q_3 are the quartile and third quartile of the weights in the minimum spanning tree respectively [1]. The edges which are higher than the threshold value are removed from the MST and the clusters are obtained. Figure 1 depicts the MapReduce model of Multi Colony ant system used in clustering the gene expression data.

The implementation of gene expression data clustering using Multi colony ant system in MapReduce framework can be outlined as follows:

1. Input the gene expression data and obtain the construction graph based on the Euclidean distance between two genes.
2. For t=1 to t_{max} do
    a) Function map<key,value>
       key: the row number which acts as ant
       value: values in the row of the construction graph
        i. In the map phase the ants choose the next node to visit based on the highest probability of the edges constructed using (1).
        ii. Ensure no cycles are formed.
       Output<key,value>
       key: the cost of the MST

value: the MST formed

b) Function reduce<key,value>
key: the cost of the MST
value: the MST formed
i. finds the best-so-far MST
ii. perform pheromone evaporation using (2) and pheromone deposition using (3)
iii. updates the probability matrix using (1).
Output<key,value>
key: the cost of the best-so-far MST
value: the best-so-far MST

3. Calculate the threshold value from the edges of the best-so-far MST found using (4).
4. Remove the edges whose value is greater than the threshold value.
5. Output the clusters formed.

The number of iterations of the multi colony ant algorithm can be predefined or the process can continue till there is no change in the best-so-far MST.

IV. Experimental Setup

The main idea behind gene expression clustering is to check whether the genes of same function are grouped in the same cluster. To verify the above gene expression data of Escherichia Coli, Yeast cell cycle were considered. A Hadoop cluster of 8 machines has been set up. The minimum spanning tree of the gene expression data were obtained using MapReduce model of multi colony ant system.

V. Result analysis

The MapReduce model of multi colony ant system was run multiple times using different number of processors (1, 2, 4 and 8) and for different data size (15.76 KB, 4.2 MB, 17.8 MB). Table1gives the execution time of MapReduce model of multi colony ant system for different file sizes and different number of processors.

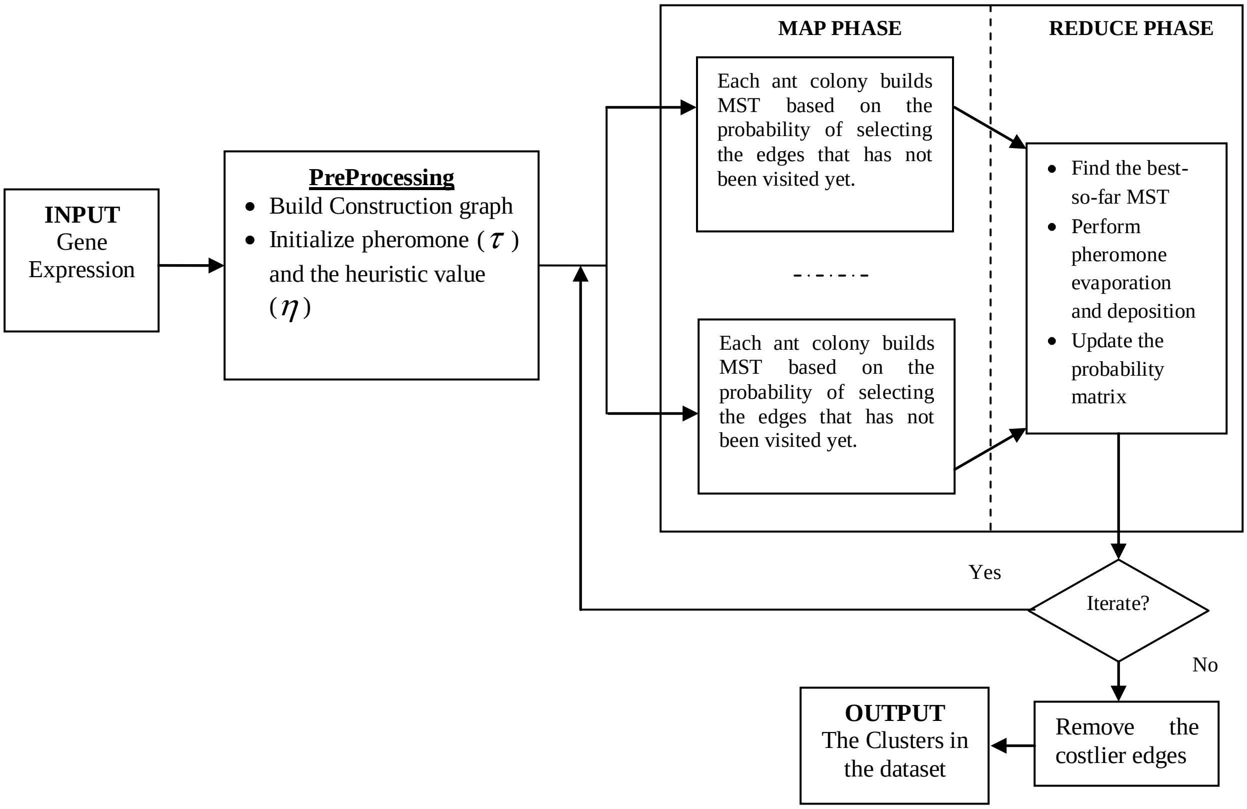
INPUT
Gene Expression
PreProcessing
Build Construction graph
Initialize pheromone (τ) and the heuristic value (η)
MAP PHASE
REDUCE PHASE
Each ant colony builds MST based on the probability of selecting the edges that has not been visited yet.
Each ant colony builds MST based on the probability of selecting the edges that has not been visited yet.
Find the best-so-far MST
Perform pheromone evaporation and deposition
Update the probability matrix
Yes
Iterate?
No
Remove the costlier edges
OUTPUT
The Clusters in the dataset

Figure 1. Flowchart of the proposed model

TABLE I. Execution Time (in seconds) of MapReduce Model of Multi Colony Ant System on File Size of Input Data vs. Number of Processors

Filesize	Number of Processors			
	1	*2*	*3*	*4*
15.76 KB	493	385	404	410
4.2 MB	936	467	351	346
17.8MB	1673	878	468	378

From the graph(Figure 2) it can be seen that for smaller file size, the execution time is almost the same when run on multiple processors. It is also observed that the execution time reduces with increase in file size. The results show that the parallel implementation of multi colony ant system using MapReduce paradigm exhibits good scalability. Many different indices for measuring the agreement between two partitions in clustering analysis with different number of clusters, exists in the literature. Adjusted rand index is used in this paper to measure the agreement between the external criteria and clustering results. The adjusted Rand index of the proposed approach is 0.947 which is much higher when compared to k-means clustering algorithm (0.563) and average-link clustering algorithm (0.572) [20].

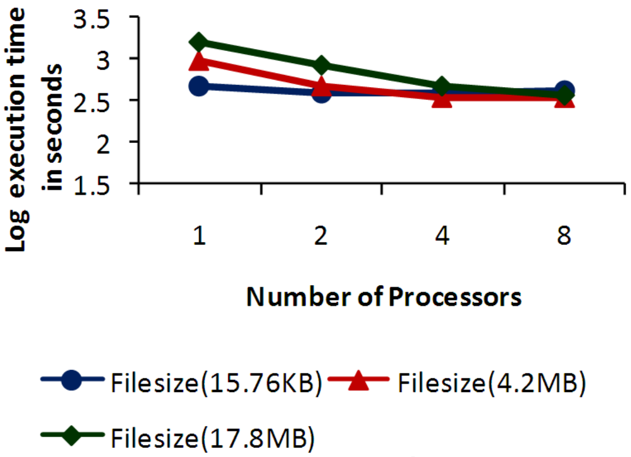

Figure2. Run Time Analysis.

VI. Conclusion

In this paper, a novel method for gene expression data clustering using MapReduce model of multi colony ant system is proposed. The system builds a minimum spanning tree using MapReduce model of multi colony ant system and finally the costlier edges in the minimum spanning tree is removed to form clusters in the given data. In order to test its scalability, the parallel algorithm was run multiple times over different number of processors. The results show that the parallel implementation of multi colony ant system using MapReduce paradigm exhibits good scalability.

References

[1] Jiawei Han and Micheline Kamber, 2006. “Data Mining: Concepts and Techniques”, *Second Edition, Morgan Kaufmann Publishers*.

[2] Tom White, 2009. “Hadoop: The Definitive Guide”, *O’Reilly Media*.

[3] Chu, S., Roddick, J.F., Su, C., and Pan, J., 2004. “Constrained ant colony optimization for data clustering”, *Trends in Artificial Intelligence: Proceedings of the 8th Pacific Rim International Conference on Artificial Intelligence*, pp. 534-543.

[4] T. Piatrik and E. Izquierdo, 2008. "An Application of Ant Colony Optimization to Image Clustering," *Proceedings of K-Space Jamboree Workshop*.

[5] Sheng-chai Chi and Chih Chieh Yang, 2008. “A Two-stage Clustering Method Combining Ant Colony SOM and K-means”, *Journal of Information Science and Engineering*, vol. 24, no. 5, pp. 1445-1460.

[6] Taher Niknam and Babak Amiri, 2010. “An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis”, *Journal of Applied Soft Computing*, vol. 10, no. 1, doi:10.1016/j.asoc.2009.07.001.

[7] Cheng-Fa Tsai, Chun-Wei Tsai, Han-Chang Wu and Tzer Yang, 2004. “ACODF: a novel data clustering approach for data mining in large databases”, *Journal of Systems and Software*, pp. 133-145.

[8] Sun Xu, Zhang Bing, Yang Lina, Li Shanshan and Gao Lianru, 2010. “Hyperspectal image clustering using ant colony optimization(ACO) improved by K-means algorithm”, *Proceedings of International Conference on Advanced Computer Theory and Engineering (ICACTE-2010)*, pp. V2-474 - V2-478.

[9] Lukasz Machnik, 2006. “Documents clustering method based on Ants Algorithms”, *Proceedings of International Multiconference on Computer Science and Information Technology*, pp. 123 – 130.

[10] P. Delisle, M. Krajecki, M. Gravel, and C. Gagne, 2001 “Parallel implementation of an ant colony optimization metaheuristic with OpenMP,” *Proceedings of International Conference on Parallel Architectures and Compilation Techniques, 3rd European Workshop on OpenMP (EWOMP’01), 2001*. Available online: http://wwwens.uqac.ca/~pdelisle/fichiers/DelisleEWOMP01Final.pdf.

[11] D.Chu, M. Till and A. Zomaya, 2010. “Parallel Ant Colony Optimization for 3D Protein Structure

Prediction using the HP Lattice Model", *Springer-Verlag Berlin, Heidelberg*, pp. 249-256.

[12] Jose M. Cecilia, Jose M. García, Andy Nisbet, Martyn Amos and Manuel Ujaldon, 2012. "Enhancing data parallelism for Ant Colony Optimization on GPUs", *Journal of Parallel and Distributed Computing*, doi:10.1016/j.jpdc.2012.01.002

[13] Jie Xiong, Xiaohong Meng and Caiyun Liu, 2010. "An improved parallel ant colony optimization based on message passing interface", *Proceedings of the First international conference on Advances in Swarm Intelligence (ICSI'10) - Volume Part I*

[14] Bihan Wu, Gang Wu, and Mengdong Yang, 2012. "A MapReduce based Ant Colony Optimization approach to combinatorial optimization problems", *Proceeding of International Conference on Computing, Networking and Communications (ICNC 2012)*, pp. 728-732.

[15] Qing Tan, Qing He and Zhongzhi Shi, 2012. "Parallel Max-Min Ant System Using MapReduce", *Advances in Swarm Intelligence Lecture Notes in Computer Science*, vol 7331, pp. 182-189 .

[16] Bhavani, R, Sadasivam, G.S. and Kumaran, R, 2011. "A novel parallel hybrid K-means-DE-ACO clustering approach for genomic clustering using MapReduce", *Proceedings of World Congress on Information and Communication Technologies (WICT-2011)*, pp. 132–137.

[17] Marco Dorigo and Thomas Stutzle, 2004. "Ant Colony Optimization", *MIT Press*.

[18] Frank Neumann and Carsten Witt, 2010. "Ant Colony Optimization and the minimum spanning tree problem", *Journal of Theoretical Computer Science*, vol-411, no. 25, pp. 2406-2413.

[19] Ka Yee Yeung and Walter L. Ruzzo, 2001. "Principal Component Analysis for clustering gene expression data", *Bioinformatics*, vol 17, no. 9, pp. 763-774.

[20] Ka Yee Yeung and Walter L. Ruzzo, 2001. "An empirical study on Principal Component Analysis for clustering gene expression data", *Technical Report, University of Washington.*